GÜNTHER GÖRZ (UNIVERSITÄT ERLANGEN-NÜRNBERG, DEPARTMENT INFORMATIK UND MAX-PLANCK-INSTITUT FÜR WISSENSCHAFTSGESCHICHTE, BERLIN)

# WissKI: Semantische Annotation, Wissensverarbeitung und Wissenschaftskommunikation in einer virtuellen Forschungsumgebung

## Zusammenfassung

Mit dem WissKI-System wird die Konzeption für eine virtuelle Forschungsumgebung vorgestellt, die aus Anforderungen an die kooperative Forschung im Bereich des Kulturerbes und seiner Dokumentation im digitalen Medium entstand und die wir im Rahmen des DFG-geförderten Projekts WissKI (›Wissenschaftliche Kommunikations-Infrastruktur‹, http://www.wiss-ki.eu/) umsetzen. Es geht dabei nicht nur um die einfache Bereitstellung und offene Verfügbarkeit von Quellmaterialien – strukturierte Texte, Grafiken, Bilder, Video, Audio – und Metadaten in digitaler Form, sondern um eine Infrastruktur für interaktives und vernetztes Arbeiten auf der Basis semantischer Tiefenerschließung. Eine Schlüsselrolle kommt hierbei dem ›Conceptual Reference Model‹ von ICOM-CIDOC als formaler Referenzontologie zu, die um geeignete Anwendungsontologien erweitert werden kann. Ihre Implementation in der ›Web Ontology Language‹ bildet in Verbindung mit verschiedenen Werkzeugen des ›Semantic Web‹ den Kern des WissKI-Systems.

## 1 Wissenschaftliche Kooperation im Internet

<1>

Wissenschaftliche Kooperation und Kollaboration im Medium des Internet ist spätestens mit dem Beginn des 21. Jahrhunderts selbstverständlich geworden. Global verteiltes Arbeiten an Daten, Bildern und Texten, basierend auf kontinuierlich wachsenden Ressourcen, z.B. in der Form digitaler Bibliotheken und Repositorien, und unterstützt durch Prozesse wie Dateitransfer, E-Mail, Videokonferenzen, etc., gehören zum Alltag der Forschung in vielen Wissenschaften. Die unter dem Titel *Semantic Web*[1] entwickelten Repräsentationsschemata und Techniken sollen dazu beitragen, die Daten in einheitlicher

Form darzustellen, semantisch zu annotieren und zu verarbeiten. Schließlich wurde mit dem Slogan ›Web 2.0‹ eine explizit als sozial bezeichnete Ebene in das World Wide Web eingeführt, mit der Web-Anwendungen, interaktives ›information sharing‹, Interoperabilität, anwender-orientierter Entwurf und Kollaboration im World Wide Web besonders unterstützt werden. Hierzu gehören beispielsweise web-basierte Gemeinschaften und Portale, die die soziale Vernetzung besonders unterstützen, sowie allgemein verfügbare Dienste und Anwendungen, Audio- und Video-Repositorien, Wikis und Blogs.

<2>

Solche medial unterstützten Formen der Zusammenarbeit, so eine weit verbreitete Ansicht, brächten ein erhebliches Potential für neue Vorgehensweisen und Denkstile in der Forschung mit sich und würden damit auch eine Grundlage schaffen für entscheidende Innovationen in der wissenschaftlichen Arbeit. Zwar kann die soziale Vernetzung im Sinne des ›Web 2.0‹ forschungsgruppenspezifisch genutzt werden, sie genügt aber dennoch nicht einer Reihe von Anforderungen, die für die Wissenschaften essentiell sind: Authentizität und Qualität der Informationen, Persistenz der Daten und Zugriffswege, Einheitlichkeit der Annotationen durch semantische Fundierung und nicht zuletzt ›Open Access‹, der freie Zugang zu wissenschaftlichen Informationen.

<3>

Im Folgenden soll am Beispiel des Kulturerbes und seiner Dokumentation im digitalen Medium ein Ansatz vorgestellt werden, den wir im Rahmen des DFG-geförderten Projekts WissKI (›Wissenschaftliche Kommunikations-Infrastruktur‹, http://www.wiss-ki.eu/) entwickeln und mit dem wir versuchen, rationale Lösungen für diese Herausforderungen zu bieten. Es geht dabei nicht nur um die einfache Bereitstellung und offene Verfügbarkeit von Quellmaterialien – strukturierte Texte, Grafiken, Bilder, Video, Audio – und Metadaten in digitaler Form, sondern um eine Infrastruktur für interaktives und vernetztes Arbeiten auf der Basis semantischer Tiefenerschließung. Auch die Unterstützung der wissenschaftlichen Kommunikation und neuer Formen der Publikation wird einbezogen.

<4>

Was die semantische Basis betrifft, werden mithilfe formaler logischer Darstellungsformen Verbindungen zwischen Wissensrepräsentationen hergestellt, so dass die Techniken des Semantischen Web[2] erweiternd nun eine neue, epistemische Ebene geschaffen werden kann, gleichsam als Web-Erweiterung der vor 30 Jahren von Brachman[3] geforderten epistemischen Abstraktionsebene für Wissensrepräsentationssprachen. Die Vernetzung wird

zum Bedeutungsträger, so dass man von einer sinnrelationalen Semantik sprechen kann, jedoch bedürfen die Etikettierungen der Kanten und Knoten des Netzwerks zuvor einer semantischen Fundierung. Für derartige semantische Modellierungen werden sog. »formale Ontologien«[4] eingesetzt, in denen Hierarchien von Begriffen und Eigenschaften in einer standardisierten (logischen) Sprache – im Semantischen Web in der Web Ontology Language OWL, zumindest aber in RDF(S),[5] – definiert werden. Diese sollten persistent im Web verankert sein, können aber auch als Modelle den Daten beigegeben werden. Auf der formalen Ebene betrachten wir die Semantik als streng kompositional im Fregeschen Sinn: D.h., die Inhaltswörter – eben als ›Etikettierungen‹ bezeichnet –, die die Begriffe und Eigenschaften repräsentieren, werden mithilfe grammatischer Regeln zu Ausdrücken zusammengesetzt, deren Bedeutung sich systematisch aus denen ihrer Teile ergibt. Für die Verarbeitung formaler Ontologien stehen dann leistungsfähige automatische Beweisverfahren zur Verfügung, die auch komplizierte, logisch zusammengesetzte Anfragen an die Datenbasen auswerten. Mittels Deduktion können Fragen nach Begründungen beantwortet und damit auch eine Unterstützung bei der Bildung wissenschaftlicher Hypothesen geleistet werden. Formale Ontologien wurden für zahlreiche Fachgebiete entwickelt und viele davon sind auch frei verfügbar.

## 2 WissKI: Ein Rahmensystem für wissenschaftliche Dokumentation und Kommunikation

<5>

Die Erstellung und Nutzung der wissenschaftlichen Objektdokumentation spielt eine zentrale Rolle in der Forschung der Institutionen des kulturellen Gedächtnisses – Museen, Archive, Bibliotheken –, die in systematischer Weise Wissen sammeln, erschließen, speichern, verwalten und kommunizieren. Die systematische Dokumentation von Sammlungsobjekten und Archivalien, von Experimenten und Messergebnissen ist Ausgangspunkt der Forschung in vielen geistes- und naturwissenschaftlichen Disziplinen; unsere ersten Anwendungsdisziplinen sind die Kunst- und Kulturgeschichte, die Archäologie und die Biologie. Darauf aufbauend zielen wir langfristig auf eine Unterstützung kooperativ organisierter Forschungsprozesse, insbesondere der Modellierung komplexer Systeme mithilfe föderierter Daten.

<6>

Ein genauerer Blick auf die gängigen Formen der Museumsdokumentation zeigt eine Vielfalt von Dokumenttypen: In der Papierform Akquisitions- und Inventarlisten und Karteikarten, Photographien, Zeichnungen, Restaurierungsprototolle, Ausstellungskataloge, Monographien, etc. In den letzten Jahrzehnten kamen Datensätze in (relationalen) Datenbanken und digitale Dokumente in Content-Management–Systemen, oft in proprietären Formaten, hinzu. Die Diskussion um Standards für Daten und Metadaten reicht nicht viel weiter als in die 1990er Jahre zurück. Was die Standardisierung von Personen- und Ortsnamen sowie von Fachterminologien betrifft, sind inzwischen sog. ›Authority Files‹ u.a. aus dem Bibliothekswesen, von der Getty Foundation, u.a. verfügbar.

<7>

Um solche Daten systematisch zu erfassen und semantisch zu erschließen, bedarf es zunächst eines methodisch aufgebauten Begriffssystems, mit dessen Hilfe sie geordnet und klassifiziert werden können, also des Fundaments jeder Theorie für das betreffende Fachgebiet. Die Modellierungsfrage ist an erster Stelle ein epistemisches und wissenschaftstheoretisches Problem: Es geht um rationale Rekonstruktion und damit um Präzisierung, Abstraktion und Formalisierung. Ein solches System von Fachbegriffen, üblicherweise als Hierarchie von Konzepten und Eigenschaften mit fachbedingten Einschränkungen formuliert, wird auch ›formale Ontologie‹ genannt. Um sie auch informationstechnisch nutzen zu können, ist zu überlegen, in welcher formalen Sprache mit welchen Zielen sie implementiert werden soll, d.h. z.B., welche Schlussfolgerungen und Anfragen eine Rolle spielen, oder ob man Daten nur semantisch etikettieren will. Die Entscheidung für ein geeignetes Notationssystem wird heute immer zum Einsatz einer Variante einer logischen Sprache führen. Die Geschichte der Wissensrepräsentationssprachen hat zwei Arten von Notationssystemen hervorgebracht: entweder solche mit unklarer Semantik oder solche, die eine semantische Fundierung besitzen, welche auf die Semantik der formalen Logik zurückgeführt wird. Zu letzteren gehören die sog. Beschreibungslogiken,[6] wovon eine relativ ausdrucksstarke Variante in der Form der erwähnten OWL-DL sich als ›Ontologiesprache‹ anbietet.

<8>

Ein wichtiges praktisches Ziel ist die Interoperabilität; Daten eines Sammlungsbestandes gewinnen umso mehr an Wert, je mehr sie mit anderen Daten verknüpft werden. Nachdem Versuche, die in Museen eingesetzten oft sehr unterschiedlichen Datenbankschemata

aufeinander abzubilden, im Großen und Ganzen misslungen waren, hat ICOM-CIDOC (Sektion für Museumsdokumentation des Internationalen Komitees der Museen) die Erarbeitung des »Conceptual Reference Model« (CRM, ISO-Standard 21127)[7] veranlasst, einer Referenzontologie für die Dokumentation des Kulturerbes, in die hinein die verschiedenen Datenbankschemata abgebildet werden können, um so Transformationsprozesse und Interoperabilität zu unterstützen. Damit wurde auch ein Paradigmenwechsel vollzogen, denn das CRM ist in innovativer Weise ereignisbasiert konstruiert, so dass die üblicherweise in Datensätzen erfassten Eigenschaften wie z.B. Urheber, Titel, Zeit, Ort, Inventarnummer, etc. nun durch ihre Rolle in bestimmten Ereignissen, wie etwa einem Herstellungs- oder Akquisitionsereignis, zusammengebunden werden. Dies bedeutet in der Tat einen konzeptionellen Abschied von der konventionellen Dokumentationspraxis: Ereignisse bilden die Klammer zwischen Handlungsträgern, begrifflichen und physischen Gegenständen, die zu bestimmten Zeiten und Orten in zeitgebundenen Objekten resultieren, welche dann durch Benennungen und Fachterminologien näher bestimmt werden können.

<9>

Neben den technischen Voraussetzungen der globalen Vernetzung und den formalen durch die vom World Wide Web Consortium (W3C) propagierten Vorschläge einer auf XML basierten Hierarchie von Repräsentationssprachen im Semantischen Web (RDF, OWL) waren die durch Referenzontologien wie das CRM geprägten Modellierungsprozesse eine entscheidende Voraussetzung für die Konzeption von WissKI.

<10>

In dem skizzierten methodischen und technischen Kontext untersucht WissKI exemplarisch die Vorgehensweisen des Sammelns, Erschließens, Speicherns, Verwaltens und Kommunizierens von Wissensbeständen in Institutionen des kulturellen Gedächtnisses und erstellt eine Plattform, die dabei mit der Besonderheit einer einheitlichen semantischen Fundierung optimale Unterstützung bieten kann und die o.g. Ziele der Authentizität und Qualitätssicherung der Informationen sowie der Persistenz der Daten und Zugriffswege garantiert. Es geht also nicht nur um die einfache Bereitstellung und offene Verfügbarkeit von Quellmaterialien – strukturierte Texte, Grafiken, Bilder, Video, Audio – und Metadaten in digitaler Form, sondern um eine Infrastruktur für interaktives und vernetztes Arbeiten auf der Basis semantischer Tiefenerschließung, die auch Unterstützung der wissenschaftlichen Kommunikation und neuer Formen der Publikation einbezieht. Wichtige Schritte dabei sind u.a. der Vergleich von Wissenselementen – Quellen und ihren Beschreibungen –, das

Auffinden von Beziehungen zwischen solchen, die themenzentrierte Suche, auch über Kollektionen hinweg, aber auch die (nicht zielgerichtete) Exploration, verschiedene Arten der Auswertung, Visualisierung und Interpretation, und nicht zuletzt die konsistente Synthese neuer Wissenselemente.

<11>

Im Zentrum der semantischen Verarbeitung steht dabei unsere Implementation des CRM in OWL-DL, das Erlangen CRM / OWL[8] (s.a. http://erlangen-crm.org/).

Für den WissKI-Prototyp wurden drei Anwendungsfälle ausgewählt:

1. die Dokumentation der Objekte in der Dauerausstellung »Renaissance – Barock – Aufklärung« im Germanischen Nationalmuseum (GNM) Nürnberg aus dem GNM-DMS (Dokumenten Management System);

2. die Projektdaten und Kommunikationsprozesse des laufenden Forschungsprojekts »Der frühe Dürer« im GNM;

3. die Biodat-Datenbank mit dem taxonomischen Material der Insektensammlung und die Expeditionstagebücher des Entomologen Wilhelm Aerts des Zoologischen Forschungsmuseums Alexander Koenig (ZFMK) Bonn.

<12>

Die Objektbeschreibungen liegen in einer semistrukturierten Form vor; teilweise sind sie in relationalen Datenbanken gespeichert. In jedem Fall sind aber auch Abschnitte mit frei formulierten Texten enthalten, die Hintergrundinformationen über Personen, Objekte, Orte, Zeiten, Materialien, Techniken, Stile, etc. enthalten und die sich auf Felder desselben oder anderer Datensätze beziehen.

<13>

Wie sollen nun in WissKI die Daten bearbeitet, aufbereitet und verknüpft werden? Der erste Schritt der semantischen Erschließung besteht darin, die Attribute der jeweiligen Beschreibungssysteme bzw. Datenbankschemata auf die Konzepte und Eigenschaften des CRM abzubilden. Da erstere oft fachspezifische Detaillierungen aufweisen, ist es in vielen Fällen sinnvoll, dafür eine sog. Bereichs- oder Domänenontologie zu definieren, die mit dem CRM verknüpft wird. Auf diese Weise können unterschiedliche Datenbestände einheitlich erschlossen werden. Mithilfe der generischen Kategorien des CRM wird eine erste Interpretationsebene eingeführt, die das Vernetzen von Wissen möglich macht. OWL-DL als

logische Sprache gestattet Verknüpfungen auszudrücken, die von einem Beweiser ausgewertet werden können, so dass auf diese Weise die Ableitung implizit dargestellten Wissens in expliziter Form möglich wird.

<14>

Techniken der Sprachverarbeitung spielen eine Schlüsselrolle für die Formalisierung der in den Freitexten ausgedrückten Inhalte. Daher müssen linguistische Hilfsmittel eingebunden werden: Wörterbücher, Programme zur morphologischen und partiellen syntaktischen Analyse und zur Erstellung semantischer Repräsentationen auf der Basis der eingesetzten formalen Ontologien. An erster Stelle sind Eigennamen von Personen, Organisationen und Orten, Zeitangaben und Fachtermini zu erkennen; dies ist ein erster Schritt zur Erkennung dargestellter Ereignisse und Ereignisfolgen. Ziel ist die Generierung semantischer Repräsentationen aus den Texten, die sich nicht unterscheiden von solchen, die direkt aus den strukturierten Datenfeldern abgeleitet werden können, z.B. für ein Herstellungsereignis: ein bestimmtes Kunstwerk wurde von Künstler N am Ort O zur Zeit Z fertiggestellt. Auch wenn wir uns mit den ersten Anwendungen fast nur im Medium der deutschen Sprache bewegen, ist grundsätzlich die Möglichkeit der Multilingualität und der Verarbeitung anderer Schriftsysteme ein Desiderat. Als Schnittstelle hierzu dient ein browser-basierter Editor, der interaktive Unterstützung bei Korrekturen der vorgeschlagenen Annotationen und auch beim Schreiben neuer Texte bietet. Darauf werden wir im nächsten Abschnitt zurückkommen.

<15>

Welche Formen des wissenschaftlichen Arbeitens sollen von einer virtuellen Forschungsumgebung wie dem WissKi-System zunächst unterstützt werden? Durch die semantische Fundierung liegt eine besondere Stärke in der Unterstützung der Wissensmodellierung, so dass Präzisierungen der Fachterminologie und Verbesserung der Genauigkeit von Objektbeschreibungen unterstützt werden. Die Hypothesenbildung wird durch die Möglichkeiten der semantischen Verknüpfungen und deren automatische Auswertung ebenso erleichtert wie die Überprüfung von Hypothesen. Hierbei ist insbesondere an solche Szenarien zu denken, in denen Erklärungen mehrere Einflussgrößen und deren Interaktionen berücksichtigen müssen. Zudem wird die inhaltliche Suche wesentlich leistungsfähiger: beispielsweise können bei logisch zusammengesetzten Anfragen auch Unter- und Oberbegriffe einbezogen werden, es können Anfragen mit Unterspezifikation bearbeitet werden und Beschreibungen verschiedener Objekte, die derselben Klasse zugehören, müssen nicht alle dieselben Attribute haben.

<16>

Mit der Möglichkeit einer einheitlichen semantischen Modellierung von Daten verschiedener Fachgebiete eröffnen sich auch Perspektiven für die Bearbeitung hochkomplexer Fragestellungen, die ein einzelnes Fachgebiet nicht zu leisten vermag. Hierzu gehört etwa die Erforschung der Biodiversität, der mittelalterlichen Stadt in allen Facetten, der Globalisierung des Wissens oder des Kulturtransfers. Unter der Bezeichnung »Transdisziplinarität« [9] wird von Mittelstraß eine neue problemorientierte kooperative Forschungsform gefordert, die u.a. unter Rückgriff auf die bedeutungskonstituierenden Begründungsverfahren in den Wissenschaftssprachen disziplinäre Grenzen überwindet. Der globale Wandel erfordert eine problemorientierte Wissenschaft, die ihre Objekte in vielfältigen Kontexten untersucht. Dabei kommt dem CRM insofern eine wegweisende Funktion zu, weil es ein wichtiges (formal-) sprachliches Mittel ist, um die geforderte praktische Einheit der Wissenschaftssprachen in disziplinübergreifenden, bedeutungskonstituierenden Prozeduren der Begründung und Rechtfertigung abzubilden.

## 3 Zur Architektur des WissKI-Systems

<17>

Als zentrale Aufgabe einer virtuellen Forschungsumgebung wurde bei der Konzeption des WissKI-Systems die Forschungsunterstützung mit semantischen Techniken und damit Werkzeugen des Semantic Web identifiziert. Die Auswahl eines Rahmensystems für WissKI führte auf das weit verbreitete Open-Source Content-Management-System Drupal (http://drupal.org/), nicht zuletzt deshalb, weil es neben einer Reihe mitgelieferter, für unsere Zwecke nützlicher Module auch ohne größere Probleme um Komponenten für die semantischen Verarbeitungsebenen, also Werkzeuge für Wissensrepräsentation und -verarbeitung, erweitert werden kann. Drupal selbst stellt Hilfsmittel für die Kommunikation – Blog, Wiki, Forum – und die Datenhaltung für Bilder und strukturierte Texte sowie eine Reihe von Werkzeugen, beispielsweise zur Navigation oder zur Formulargenerierung bereit. Die Kommunikationswerkzeuge werden in WissKI nur in moderierter Form angeboten, d.h., dass stets ein Kurator die Qualitätskontrolle durchführt, bevor Informationen publiziert werden.

<18>

Die WissKI-spezifischen Erweiterungen, die die semantische Basis des Systems bilden, lassen sich als Abfolge von Schichten darstellen (s. Abb. 1): An oberster Stelle steht die Erlanger Implementation des CRM in OWL-DL. Sie wird erweitert durch einige Klassen und Eigenschaften, darunter auch Basis-Datentypen, die sich für die praktische Arbeit als nützlich erwiesen haben und die unter der Bezeichnung ›WissKI-Basis-Ontologie‹ zusammengefasst werden. Die verschiedenen Domänenontologien, z.B. Biodat, die Gemälde- und die Dürer-Ontologie, sind gleichberechtigt darunter gestellt, so dass deren Konzepte Unterkonzepte von CRM bzw. der Basis-Ontologie sind. Die importierten Daten werden in Instanzen der jeweiligen Ontologie überführt; sämtliche Instanzen werden serialisiert als Elementaraussagen ›Subjekt – Prädikat – Objekt‹ in der Form von sog. RDF-Tripeln abgelegt, wofür ein eigener besonders effizienter Tripelspeicher (ARC, s. http://arc.semsol.org/) bereit steht. Auf der nächsten Ebene befinden sich die aus globalen sog. ›Authority Files‹ importierten Normdaten-Lexika für Personen- und Ortsnamen und Thesauri der Fachterminologien.

Sie sind sämtlich in dem XML-basierten Format SKOS (»Simple Knowledge Organization System«) [10] gespeichert, wofür eigens ein Konvertierungsprogramm geschaffen wurde, sodass z.B. die lizenzpflichtigen Getty-Thesauri (Geographic Names, United List of Artist Names, Art and Architecture) oder die Personennamendatei der Nationalbibliothek, Geonames (http://geonames.org), Graesses Orbis Latinus, und andere einheitlich dargestellt sind. Für jede dieser Schichten gibt es Programmierschnittstellen (APIs) zum Daten-Import und -Export; administrative Metadaten werden durch das »Open Archive Initiative Protocol for Metadata Harvesting« (OAI-PMH) bereitgestellt, das wie ein Umschlag um die abgefragten (Meta-) Daten funktioniert.

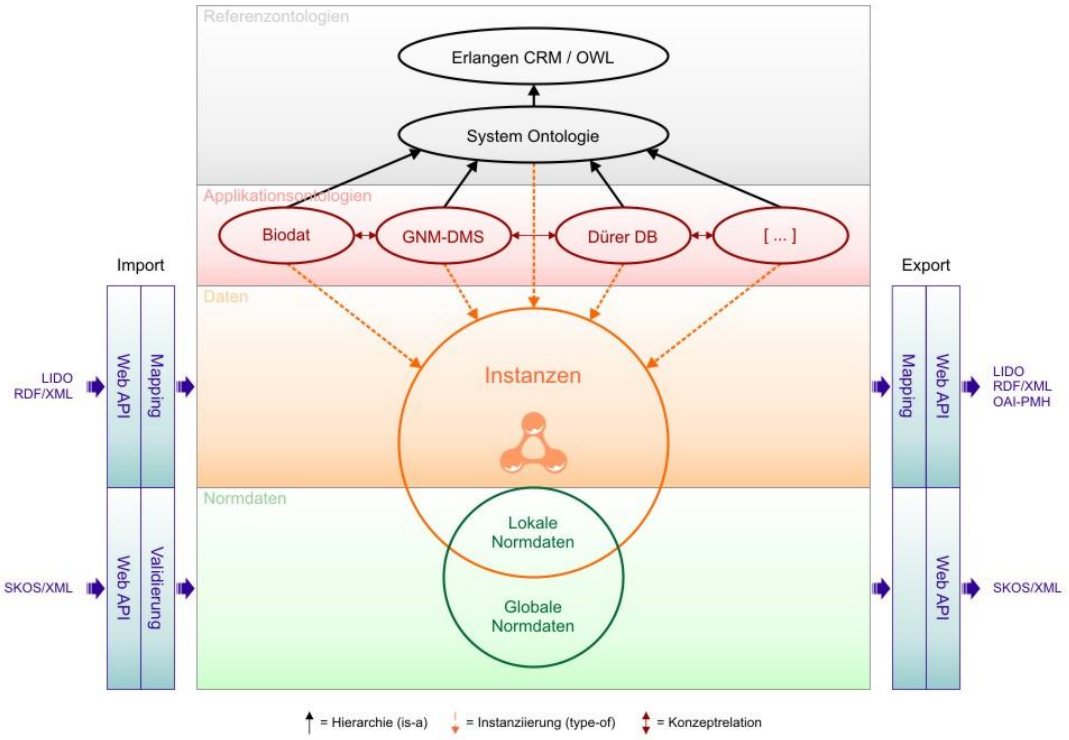


Abb.1: Formale Ontologien, Instanzen und Normdaten im WissKI-System

<19>

Die semantische Basis ist der Schlüssel für das Erzeugen und Publizieren einheitlicher normierter Annotationen zu den Beschreibungen der Objekte der wissenschaftlichen Sammlungen. Wie werden solche Annotationen erzeugt und verwaltet? Die in den Ausgangsdaten in strukturierten Feldern direkt gegebenen Eigenschaftsausprägungen werden beim Import in das WissKI-System mittelbar oder ggf. direkt auf die CRM-Konzepte und -Eigenschaften abgebildet. Für die Freitexte wurden spezielle Parser zur Erkennung von Namen und Zeitangaben implementiert, die die lokalen Lexika und ein eigenes morphologisches Modul zur Bestimmung der Grundformen benutzen und die semantische Repräsentationen in der Form von RDF-Tripeln erzeugen.[11] Als Schnittstelle zu den Parsern wurde ein browserfähiger Editor (TinyMCE) integriert – der für alle Eingabemodalitäten von Drupal, z.B. Wikis, konfiguriert werden kann –, sodass damit auch Annotationen korrigiert und beim Schreiben von Texten automatisch Annotationen erzeugt werden können (siehe Abb. 2). Mit den Namen, Zeitangaben und Fachtermini sind bereits wichtige Bausteine für Ereignisbeschreibungen vorhanden. Der nächste schwierige Schritt besteht in der Erkennung des Ereignis- oder Handlungstyps, der durch Verben bzw., da in wissenschaftlichen Texten oft der Nominalstil vorherrscht, in Verb-Substantiv-Kombinationen ausgedrückt ist. Für diese Wörter benötigt der Parser ein Valenzlexikon, in welchem dem

jeweiligen Ereignisprädikat auch die thematischen Rollen wie Handlungsträger, direktes und indirektes Objekt, Ort und Zeit, etc. zugeordnet sind.

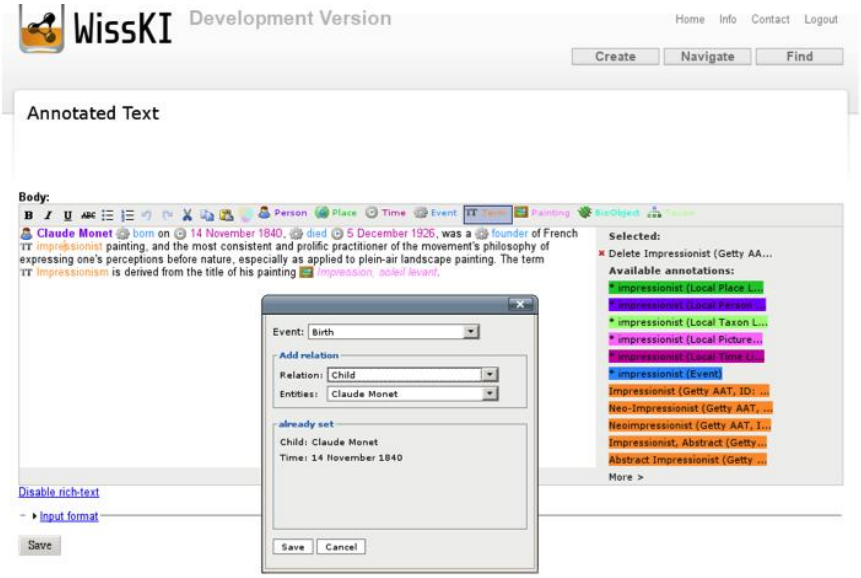


Abb.2: Freitext-Annotation

<20>

Was die semantischen Repräsentationen betrifft, greifen wir auf die Diskursrepräsentationstheorie zurück,[12] um auch satzübergreifende Abhängigkeiten erfassen zu können. Ereignisbeschreibungen selbst werden in einer von einem Vorschlag von Davidson [13] abgeleiteten logischen Darstellung formuliert, die direkt dem allgemein verwendeten RDF-Tripel-Format entspricht und damit voll integrierbar ist. Die nächste Herausforderung wäre dann, auch Folgen von Ereignissen zu erkennen, etwa mit Hilfe von Techniken der Planrekonstruktion, wonach Ereignisse wie Planungsoperatoren zu beschreiben wären durch Vorbedingungen und Ergebniseigenschaften, die dann wieder Vorbedingungen für Folgeereignisse wären – dies liegt aber jenseits des aktuellen Projekthorizonts.

<21>

Wie kann nun in dem dargestellten Wissensbestand nach inhaltlichen Gesichtspunkten navigiert werden? Da der typische WissKI-Anwender eher wenig vertraut mit logischen Ausdrücken und entsprechenden Anfragesprachen sein wird, werden ebenso wie für die Dateneingabe auch für Anfragen geeignete Schablonen aus den Ontologien generiert bzw.

mit ihrer Hilfe konfiguriert. Grundlage sind dafür Pfade von Konzepten, die durch Eigenschaften miteinander verbunden sind und an deren Ende immer ein Konzept steht, das durch einen konkreten Wert instanziiert wird, der in die Schablone einzutragen ist. Diese Pfade werden in einem Vorverarbeitungsschritt stückweise vom Definitions- zum Wertebereich einer Eigenschaft durch den Inferenzmechanismus bestimmt und zusammengesetzt. Sie sind zugleich semantische Navigationswege durch den Wissensbestand. Allgemein soll der Zugang zu dem repräsentierten Wissen durch typische in schematischer Form gegebene Anfragemuster[14] erleichtert werden. Anfragen werden durch den Inferenzmechanismus bearbeitet, der auch lediglich implizit repräsentiertes Wissen erschließen und explizit repräsentieren kann.

<22>

Durch die Wahl von OWL-DL als Wissensrepräsentationssprache erhält man eine ausdrucksstarke Teilsprache der Standardlogik mit dem wesentlichen Vorteil der Entscheidbarkeit. Für sie stehen effiziente Inferenzverfahren zur Verfügung, die sowohl Korrektheit (es werden keine falschen Schlüsse gezogen) als auch Vollständigkeit (es werden alle korrekten Schlüsse gezogen) garantieren. Hierzu gehören u.a. Racer (http://www.racer-systems.com/) und Pellet (http://clarkparsia.com/pellet/), zu deren Standardleistungen gehört, Konsistenz festzustellen, d.h. Widersprüche in den Daten zu finden, automatische Klassifikation zu leisten, d.h. die Konzepthierarchie komplett zu berechnen, und die Instanzen für jede Klasse (und umgekehrt) zu bestimmen. Eine ihrer Stärken liegt darin, ›normale‹ Datenbankabfragen, die in der im Semantic Web gebräuchlichen Anfragesprache SPARQL[15] mit terminologischem Schließen zu verbinden. Ein Beispiel hierfür wäre die Suche nach allen Urhebern graphischer Kunst, wobei dieses Prädikat zwar in der formalen Ontologie, jedoch nicht explizit in den Datensätzen enthalten sein möge, sondern nur dessen spezielle Ausprägungen (Unterklassen) wie Zeichnung, Holzschnitt, Kupferstich, Radierung, etc. Ein weiteres Beispiel wäre die Suche nach allen Stillleben, die Insekten darstellen, aber kein Obst – man beachte die Negation! Insbesondere sei auf die Besonderheit der sog. ›Offenen-Welt-Semantik‹ von OWL-DL hingewiesen: Anfragen nach nicht vorhandenen Sachverhalten werden nicht wie bei relationalen Datenbanksystemen mit ›falsch‹, sondern mit ›unbekannt‹ beantwortet. Es wird also keine Vollständigkeit der Datenbasis unterstellt und die Antwort auf der epistemischen Ebene gegeben.

<23>

Künftig soll es auch möglich sein, noch vielfältigeren Gebrauch von weiteren Softwarekomponenten und externen Diensten zu machen. So können Bilder bisher nur abgespeichert, aber noch nicht im System bearbeitet werden. An dieser Stelle könnte mit digilib (http://digilib.berlios.de/) ein leistungsfähiges, mit jedem Browser einsetzbares System zur Online-Bildbearbeitung für wissenschaftliche Zwecke in Client/Server-Architektur eingesetzt werden, das wir bereits in anderen Anwendungen benutzen. Neben seinen vielfältigen Möglichkeiten der Bildbearbeitung verfügt es auch über die Möglichkeit, hochaufgelöste Bilder zu annotieren sowie bearbeitete Bilder in stabilen URLs zu referenzieren, die sich besonders für die wissenschaftliche Kommunikation eignen – es müssen nicht mehr voluminöse Bilddateien übertragen werden, sondern die Bildadresse zusammen mit einer Bearbeitungsvorschrift.

<24>

Für die Arbeit mit XML-annotierten Quelltexten und die Erstellung von Publikationen in verschiedenen Formaten – in unserem Fall kommen am ehesten TEI-konforme [16] Darstellungen für digitale Editionen in Frage – gibt es mit Arboreal (http://archimedes.fas.harvard.edu/arboreal/) ein sehr vielseitiges Werkzeug. Arboreal erlaubt u.a., mit Paralleltexten zu arbeiten, Lemmatisierung durch Aufruf externer morphologischer Dienste für eine Reihe von Sprachen durchzuführen, externe Fachlexika aufzurufen, unterstützt den Aufbau von Terminologielisten und die Referenzierung von Fachtermini. Die aktuelle Version von Arboreal verfügt über einen eigenen Browser; künftig sollte auch eine Internetbrowser-fähige Version verfügbar sein, die dann ebenfalls in das WissKI-System integriert werden könnte.

## 4 Ausblick

<25>

Mit dem skizzierten Modellierungsansatz und seiner Implementation im WissKI-System in der Kombination mit qualitätskontrollierten Publikationsverfahren ist für den Bereich der Objektdokumentation ein innovativer erster Schritt gemacht. Sobald man sich aber auf reale Anwendungsszenarien einlässt, treten bald – durchaus allgemein bekannte – Probleme auf, für die in den konkreten Praxisfeldern bei Weitem noch keine Lösungen in Sicht sind. Dies sei kurz anhand der Fragen der Skalierbarkeit und der Nichtmonotonie erläutert.

<26>

Ein gewichtiges Problem, mit dem sich gegenwärtig einige große Forschungsprojekte befassen, und für das auch in WissKI eine Lösung gefunden werden muss, sind die Kapazitätsgrenzen der aktuellen Beweissysteme. Selbst in WissKI haben wir schon mit den drei genannten Szenarien Instanzendaten im Umfang von etwa 1,5 Millionen Tripeln, was signifikant über dem liegt, was moderne Inferenzmaschinen wie Racer oder Pellet verarbeiten können. Auch die bekannte Vorgehensweise, dann auf unvollständige Beweisverfahren überzugehen, verschiebt das Problem nur und kann auf Dauer keine Lösung sein, denn die Datenmengen werden weiterhin überproportional zunehmen. Das Massenproblem ist also eine gewaltige algorithmische Herausforderung. Eine Möglichkeit, die sich zur Erprobung anbietet, besteht in der Zusammenfassung der Instanzen in Äquivalenzklassen nach gewissen Kriterien, vgl. SHER.[17]

Es ist eine offene Forschungsfrage, ob es gelingt, sinnvoll große Äquivalenzklassen zu bilden und dann nur jeweils einen Repräsentanten für eine bestimmte Anfrage in die Verarbeitung einzubeziehen, um so wieder in den Leistungsbereich der genannten Inferenzmaschinen zu kommen.

<27>

Eine weitere offene Forschungsfrage ist der Umgang mit inkonsistenten Daten. Da WissKI sich mit der Modellierung des Dokumentationsprozesses beschäftigt, muss es Szenarien unterstützen, in denen einander widersprechende Dokumente zu verarbeiten sind. Gibt es beispielsweise zwei Dokumente mit unterschiedlichen Geburtsdaten für dieselbe Person, so kann damit nur gearbeitet werden, wenn die jeweiligen Rechtfertigungen, hier also Angaben über die Quellen, berücksichtigt werden. Die gegenwärtige Instanzenspeicherung unterstützt dies jedoch nicht – sie speichert nur Fakten, aber keine Abhängigkeiten. Dies ist nur eine Variante des Problems der Nichtmonotonie, das auch im Zusammenhang der Repräsentation und Verarbeitung generischer Aussagen eine Rolle spielt. Gerne macht man Allaussagen, wohl wissend, dass es Ausnahmen gibt: »Alle Vögel fliegen« ist eine bekannte Standardannahme, die man so lange aufrecht erhält, bis man einen Strauß oder Pinguin vor sich hat. Spätestens dann wird man gewahr, dass man die wissenschaftliche Argumentation nicht einstellen kann, weil eine Inkonsistenz vorliegt, sondern Verfahren braucht, die nicht an den Grenzen der (monotonen) Logik halt machen. Beide Varianten der Nichtmonotonie sind in der Wissenschaftspraxis alltäglich und theoretisch auch gut untersucht.[18] Wir gehen davon aus, dass die für das praktische Schließen mit Standardannahmen entwickelten Verfahren der Begründungsverwaltung (»Reason Maintenance«) bzw. der ›Default‹-Regeln an dieser

Stelle in die Praxis eingebracht werden können. Generell sehen wir die Frage der Nichtmonotonie nicht als logisches, sondern als epistemisches Problem, das somit auch nicht auf der Ebene der logischen Verarbeitung alleine zu lösen ist. Ein kleines Experiment für den letzteren Fall wurde erfolgreich durchgeführt.[19] Der Lösungsansatz besteht darin, die Verarbeitung in zwei Phasen zu trennen: In einer Vorverarbeitungsphase werden die Default-Regeln angewandt, die je nach vorliegender Evidenz (hier: Rabe oder Pinguin) dann eine jeweils individuelle Instanz der generischen Aussage erzeugen, sodass in der zweiten Phase monoton geschlossen werden kann. Eine generelle, praktisch einsetzbare Lösung dieses Problemkomplexes, der bei der wissenschaftlichen Hypothesenbildung bzw. ihrer Rekonstruktion eine wichtige Rolle spielt, ist jedoch noch nicht in Sicht.

<28>

Der WissKI-Ansatz kann mit seinen partiellen Lösungen auf jeden Fall dazu dienen, in realen Anwendungssituationen der Dokumentationspraxis und dem Einsatz als Instrument in der Forschung systematisch – und mit Drupal auch automatisch unterstützt – Erfahrungen zu sammeln, von deren Auswertung neue Einsichten in den Forschungsprozess und die Anforderungen an sinnvolle Hilfsmittel zu seiner Unterstützung zu gewinnen. Vielleicht gelingt es dann auch, Rekonstruktionsansätze für bestimmte reale Forschungsprozesse, wie sie beispielsweise Graßhoff[20] modelliert hat, derart zu verallgemeinern, dass sich daraus nützliche Angaben zur Konfiguration innovativer Werkzeuge für die Praxis des vernetzten wissenschaftlichen Arbeitens ableiten lassen.

<29>

Danksagung: Der Autor dankt Mark Fichtner, Georg Hohmann, Siegfried Krause, Karl-Heinz Lampe (+), und Martin Scholz für hilfreiche Hinweise. Ein herzlicher Dank gebührt der Deutschen Forschungsgemeinschaft für die Förderung des WissKI-Projekts (GO 452/6-1).

## Zum Autor

Geboren 1947 in Nürnberg; Studium der Mathematik, Informatik und Philosophie (Schwerpunkte: Logik, Sprachphilosophie, Wissenschaftstheorie und -geschichte) in Erlangen; Promotion zum Dr.-Ing. in Informatik (Sprachverarbeitung). 1972-1987 wiss. Mitarbeiter an der Universität Erlangen-Nürnberg; 1981 Visiting Assistant Professor UCLA, Los Angeles; 1985, 2004 Gastforscher am CSLI, Stanford University; 1999 Gastforscher am ICSI, UCB, Berkeley. 1989-1991 Professor für Informatik, Universität Hamburg; seit 1991

Professor für Informatik (Künstliche Intelligenz), Friedrich-Alexander-Universität Erlangen-Nürnberg; Zweitmitglied der Philosophischen Fakultät. Gastwissenschaftler am Max-Planck-Institut für Wissenschaftsgeschichte, Berlin.

Schwerpunkte: Maschinelle Sprachverarbeitung; Angewandte Logik, Wissensrepräsentation und -verarbeitung; Digitale Medien und Dokumentation des Kulturerbes. Weitere Forschungsinteressen: Logik und Sprachphilosophie, Ontologie, Wissenschaftstheorie und Wissenschaftsgeschichte, insbes. in Antike und Mittelalter.

Postadresse: Prof. Dr. Günther Görz, Univ. Erlangen-Nuernberg, Department Informatik 8/KI, Haberstrasse 2, 91058 ERLANGEN, Fon: (+49 9131) 852-8701/2

Mail: goerz@informatik.uni-erlangen.de

Web: http://www8.informatik.uni-erlangen.de/inf8/en/goerz.html

---

1 Hitzler, P.; Krötzsch, M.; Rudolph, S.; Sure, Y. (Hrsg.): Semantic Web: Grundlagen, eXamen.press, Springer, Berlin [u.a.], 2008.

2 Hitzler, P.; Krötzsch, M.; Rudolph, S.; Sure, Y. (Hrsg.): Semantic Web: Grundlagen, eXamen.press, Springer, Berlin [u.a.], 2008.

3 Brachman, R. J.: What's in a concept: structural foundations for semantic networks, International Journal of Man-Machine Studies, Bd. 9, Nr. 2, March 1977, S. 127–152.

4 Noy, N. F.; McGuinness, D. L.: Ontology Development 101: A Guide to Creating Your First Ontology, Online, 2001; Hitzler, P.; Krötzsch, M.; Rudolph, S.; Sure, Y. (Hrsg.): Semantic Web : Grundlagen, eXamen.press, Springer, Berlin [u.a.], 2008.

5 Vgl. Hitzler, P.; Krötzsch, M.; Rudolph, S.; Sure, Y. (Hrsg.): Semantic Web: Grundlagen, eXamen.press, Springer, Berlin [u.a.], 2008.

6 Hitzler, P.; Krötzsch, M.; Rudolph, S.; Sure, Y. (Hrsg.): Semantic Web: Grundlagen, eXamen.press, Springer, Berlin [u.a.], 2008, Kap. 6.

7 Vgl. Doerr, M.: The CIDOC conceptual reference model: an ontological approach to semantic interoperability of metadata, AI Magazine, Bd. 24, Nr. 3, September 2003, S. 75–92.

8 Goerz, G.; Oischinger, M.; Schiemann, B.: An Implementation of the CIDOC Conceptual Reference Model (4.2.4) in OWL-DL, in Proceedings CIDOC 2008 – The Digital Curation of Cultural Heritage. Athen, Benaki Museum, 15.–18.09.2008, ICOM CIDOC, Athen, September 2008, S. 1–14.

9 Mittelstraß, J.: Transdisciplinarity – New Structures in Science, in Innovative Structures in Basic Research. Ringberg-Symposium, 4–7 October 2000, Nr. 5 in Max Planck Forum, München, 2002, S. 43–54.

10 Miles, A.; Bechhofer, S.: SKOS Simple Knowledge Organization System, World Wide Web Consortium, Aug. 2009.

11 Goerz, G.; Scholz, M.: Content Analysis of Museum Documentation in a Transdisciplinary Perspective, in Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELTR 2009), Association for Computational Linguistics, ACL, Athens, March 2009, S. 1–9.

12 Vgl. Kamp, H.; Reyle, U. (Hrsg.): From Discourse to Logic, Kluwer, Dordrecht, 1993; Fischer, I.; Geistert, B.; Goerz, G.: Incremental Semantics Construction and Anaphora Resolution Using Lambda-DRT, in Botley, S.; Glass, J. (Hrsg.): Proceedings of DAARC-96 – Discourse Anaphora and Anaphor Resolution Colloquium, Lancaster, July 1996, S. 235–244; Bücher, K.; Goerz, G.; Ludwig, B.: Corega Tabs: Incremental Semantic Composition, in Goerz, G.; Haarslev, V.; Lutz, C.; Moeller, R. (Hrsg.): KI-2002 Workshop on Applications of Description Logics, Proceedings, Bd. 63 von CEUR Workshop Proceedings, Gesellschaft für Informatik e.V., CEUR, Aachen, September 2002.

13 Davidson, D.: Handlung und Ereignis, Theorie, Suhrkamp, Frankfurt am Main, 1985.

14 Constantopoulos, P.; Dritsou, V.; Foustoucos, E.: Developing Query Patterns, in Research and Advanced Technology for Digital Libraries, Bd. 5714 von Lecture Notes in Computer Science, Springer Verlag, Berlin etc., 2009, S. 119–124.

15 Hitzler, P.; Krötzsch, M.; Rudolph, S.; Sure, Y. (Hrsgb.): Semantic Web: Grundlagen, eXamen.press, Springer, Berlin [u.a.], 2008, Kap. 7.

16 Ide, N.; Veronis, J. (Hrsgb.): Text Encoding Initiative. Background and Context, Kluwer, Dordrecht, 1995, Also in: Computers and the Humanities. Vol. 29, No. 1–3 (1995).

17 Srinivas, K.: OWL Reasoning in the Real World: Searching for Godot, in Description Logics, 2009, Proceedings of the 22nd International Workshop on Description Logics (DL 2009), Oxford, UK, July 27-30.

18 Vgl. Brewka, G.; Dix, J.; Konolige, K.: Nonmonotonic reasoning: an overview, Nr. 73 in CSLI Lecture Notes, Center for the Study of Language and Information, Stanford, Calif., 1997.

19 Görz, G.: »Generics and Defaults«. Zum technischen Umgang mit Begriffssystemen, Standardannahmen und Ausnahmen, in Methodisches Denken im Kontext, mentis, Paderborn, 2007, S. 383–401, Festschrift für Christian Thiel zum 70. Geburtstag.

20 Graßhoff, G.; May, M.: Methodische Analyse wissenschaftlichen Entdeckens, Kognitionswissenschaft, Bd. 5, 1995, S. 51–67.